DOCUMENT RESUME

ED 121 302 IR 003 280

AUTHOR Schipma, Peter B.
TITLE Research Toward Enhancing Retrieval Effectiveness.
SPONS AGENCY National Science Foundation, Washington, D.C. Office of Science Information Services.
PUB DATE 9 Mar 76
NOTE 13p.; Paper presented at the Annual Meeting of the National Federation of Abstracting and Indexing Services (Columbus, Ohio, March 9, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS Bibliographic Coupling; Bibliographies; Computer Programs; Cost Effectiveness; Data Bases; *Indexing; *Information Retrieval; *Information Systems; Library Automation; On Line Systems; *Relevance (Information Retrieval); *Search Strategies; Thesauri; Vocabulary
IDENTIFIERS Computerized Searches; ITT Research Institute

ABSTRACT

Two programs addressed the development of information retrieval systems useful to a variety of users. One program was designed to examine aspects of indexing and information display. Experiments with different indexing and display systems revealed that for good relevance judgements, a display with full citation, keyboards, and abstract is necessary. However, the cost of searching abstracts is high, and adequate retrieval can be made using titles and keyboards. The second program was undertaken to improve computer-search quality and efficiency for large bibliographic files. An algorithm was generated whereby citations are clustered for presentation to the user. Preliminary findings indicated that a term map is required which would project found vocabulary up the hierarchy toward greater generality. (CH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED121302

RESEARCH TOWARD ENHANCING RETRIEVAL EFFECTIVENESS

PETER B. SCHIPMA, IIT RESEARCH INSTITUTE

(PRESENTED AT NFAIS ANNUAL MEETING, 9 MARCH 1976
COLUMBUS, OHIO)

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

This Research was Reported by
NSF Office of Science Information Service

IR 003 880

INTRODUCTION

IIT Research Institute has conducted several research programs under contracts and grants from the National Science Foundation Office of Science Information Services. While the earlier programs (1968-1972 period) were primarily concerned with developing viable information retrieval systems, recent work has concentrated on improvement in the quality and cost-effectiveness of retrieval. These programs have been aimed toward the discovery of information useful to a variety of systems and system operators. I shall discuss two of these projects briefly today. One has been completed, and detailed information is available in the Final Report. The other program is current - a semi-annual report is now available, and a Final Report will be distributed at the end of this year.

STUDY OF INDEXING AND INFORMATION DISPLAY

The first project was conducted from November 1973 to April 1975.

Program Goals

Studies of indexing and display have been performed for many years. They have ranged from studies of the value of content indexing as opposed to citation indexing to studies of automated indexing. Many of these failed to be definitive because no user community had available one large data bank containing all of the variables to be tested.

Using the facilities of the Computer Search Center (CSC), IIT Research Institute (IITRI) was able to overcome some of these limitations. The recently completed program was possible because of:

- the existence of large machine-readable bibliographic record files representing the same set of documents in different but related ways*

*For this program, a bibliographic record is comprised of citation information (author and location, title, source (e.g. journal), source subelements (e.g. volume, issue, pagination, date)), plus additional information such as abstracts, index terms, molecular formulae, etc.

- the existence of software capable of performing identical manipulations on each of the above,
- the existence of well-tested user questions, and
- the existence and availability of people familiar with the data base(s) and software system.

Data bases used in this study were the Chemical Abstracts Services Condensates, CASIA and CBAC - all converted to ITTRI-format for searching purposes.

Program Strategy

This was a three-experiment program designed to quantify several aspects of indexing and information display.

I. Effect of Record Completeness on Relevancy Judgement

This experiment tested the effects of completeness of records on retrieval efficiency - the same set of questions was used to search several versions of the same data base. These versions differed in the amount of material available for searching (titles only, plus index terms, etc.)

II. Effect of Indexing Methodology on Retrieval Efficiency

This experiment studied several indexing methods for their effects upon retrieval efficiency.

III. Effect of Information Display on Relevancy Judgement

This experiment tested the effects of completeness of display on relevance judgement. The total set of retrieved documents, found by all search methods, was judged for relevance. Only certain parts of the records were given to each reviewer, to determine the effects of record completeness on their judgement of relevancy.

Program Findings

Experiment #1 - Effect of Record Completeness on Relevancy Judgement

Three levels of record completeness (citation information only,

citation information plus keywords, and citation + information plus keywords and abstract) were searched.

This experiment showed that the more complete a retrieval, the more likely is its selection as a hit. The overall average retrieval for citation information was 31.4%, for citation + information plus keywords was 54.5%, and for citation information plus keywords and abstracts was 74.2%. Adding keywords + increased the likelihood of getting a hit by 23% while adding either abstracts also increased the likelihood of a hit by another + 13%. The addition of all the other elements (index terms, bibliography numbers, molecular formula, etc.) added another 24%.

While CPC costs are not absolutely congruous to citizens of others, the data in Figure 1 indicate the expected increments in cost with increase in the size and complexity of a data search. Keywords contain quite a bit in performance for little incremental cost. The searching of full abstracts provides a similar improvement in performance at a considerable increase in cost. The activity figures are normalized to $100.00.

Figure 1

	CITATION INFORMATION	CITATION INFO. + KEYWORDS	CITATION, KEYWORDS + ABSTRACT	FULL RECORD
PERFORMANCE LEVEL	31.7%	54.5%	74.2%	100%
COST	$34.00	$41.00	$74.00	$100.00

OVERALL PERFORMANCE/COST FOR SEARCHES OF FILES AT 4 VIEWS LEVELS OF RECORD COMPLETENESS (USING CPC COST FIGURES).

Experiment #2 - Effect of Indexing Methodology on a Retrieval Efficiency

This experiment related indexing methods to a retrieval efficiency. The major portion of the work was carried out on a two files. The Uncontrolled file was used to represent unstructured, uncontrolled indexing and the Table file (the view-ordered navigation of the CPC Integrated Subject File) was used to represent controlled structured indexing.

To obtain a baseline measure for appearance of terms in given data elements, both the Condescriptor and the Title files were searched for single terms. The percent found accuracy for each data element was recorded for each term. Data for this study are given in Figure 1.

The pair of both Condescriptor and Title were searched. It was shown that the citizen a term was in a statistical table. The term often is was found in Title. In an overall average, for this sample of terms the Condescriptor keywords were the more discriminatory fields.

The very important term searched Title did not replace its various data and beneficial for search purposes. This confirmed previous findings by both (1968) and the researchers of Georgia Tech's studies of the SDC Management Decision File (SDC). Which subject indexing was of benefit when searching statistical tasks. It can pointed that citation and business information for searching subject concepts.

Searches were made of the Title file for the questions previously searched against the other versions of the data base. The important terms were obvious from the results:

- many more records were extracted from the data base which had not been found via searches of the citation and business data, and
- many but of the records identified via citation and text word searches were the same as those identified via the index search.

While a total of 7,52 records had been identified by searches of all versions of the citation and business information, the searches of the index file (Title) identified 4988. But only 828 of these 4988 were common with those contained in the Title. Thus, there were actually 11,988 identified records from the use of the searches in point of fact. searches of citation and business data alone and have considerably better all (1%) than those of index data alone are (8%). —To secure complete retrieval, both are required.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]			[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

[illegible]

[illegible]

1) Tree 167
2) Bush 202

(etc.) (etc.)

8) User specifies which of the Rn subsets he wishes to obtain and has those printed.

This scenario obviates many of the problems. The synonym problem is handled explicitly by the word map. Term ambiguity may be handled by building limited associations into the map, the key questions now are:

1. Are our preliminary results of general validity?
2. Can the required file access and the computations be be done in times compatible with an on-line environment?
3. How expensive would it be to construct a functional word map?

We will continue work toward definitive answers to these questions.